Geographical trends in research: a preliminary analysis on authors' affiliations

Geographical trends in research: a preliminary analysis on authors' affiliations

Andrea Mannocci, Francesco Osborne, Enrico Motta

Knowledge Media Institute, The Open University, Milton Keynes, UK
name.surname@open.ac.uk

Abstract. In the last decade, research literature reached an enormous volume with an unprecedented current annual increase of 1.5 million new publications. As research gets ever more global and new countries and institutions, either from academia or corporate environment, start to contribute with their share, it is important to monitor this complex scenario and understand its dynamics.
We present a study on a conference proceedings dataset extracted from Springer Nature Scigraph that illustrates insightful geographical trends and highlights the unbalanced growth of competitive research institutions worldwide. Results emerged from our micro and macro analysis show that the distributions among countries of institutions and papers follow a power law, and thus very few countries keep producing most of the papers accepted by high-tier conferences. In addition, we found that the annual and overall turnover rate of the top 5, 10 and 25 countries is extremely low, suggesting a very static landscape in which new entries struggle to emerge. Finally, we highlight the presence of an increasing gap between the number of institutions initiating and overseeing research endeavours (i.e. first and last authors' affiliations) and the total number of institutions participating in research. As a consequence of our analysis, the paper also discusses our experience in working with affiliations: an utterly simple matter at first glance, that is instead revealed to be a complex research and technical challenge yet far from being solved.

Keywords: Scholarly knowledge, affiliations, conferences, scientometrics, research, scigraph

1 Introduction

Over the last decade, research started to scale up in terms of produced volume of papers, authors and contributing institutions. Nowadays, research literature is estimated to round up 100-150 million publications with an annual increase rate around 1.5 million new publications [2]. Such a complex, global-scale system is worth studying in order to understand its dynamics and internal equilibria. In particular, the study of authors' affiliations [7,15] has concrete impact on the interpretation of research as a complex phenomenon inserted in a delicate socioeconomic and geopolitical scenario.

In this study, we present an analysis on a dataset of conference proceedings metadata covering the 1996-2017 period, which was distilled from SciGraph[1], a free linked open data (LOD) dataset about scholarly knowledge published and curated by Springer Nature. In particular, we first present a *macro analysis* on the full dataset, including conference proceedings across several scientific disciplines (e.g. computer science, life sciences, chemistry, engineering) and then a *micro analysis*, which focuses on three high-tier conferences close to our area of expertise: the International Semantic Web Conference (ISWC), the Extended Semantic Web Conference (ESWC), and the International Conference on Theory and Practice of Digital Libraries (TPDL).

The main contribution of this work is threefold. Firstly, we found that, over the observed period, the distributions of institutions and papers among countries follow a power law, consistently to what previously demonstrated in the literature across the 1981-2010 period [12,10,4,15]. Therefore, very few subjects keep producing most of the papers accepted by scientific conferences. Secondly, we show how the annual and overall turnover rate of the top 5, 10 and 25 countries is extremely low, suggesting a very static landscape in which new entries struggle to emerge. Finally, we highlight an increasing gap between the number of institutions initiating and overseeing research endeavours (i.e. first and last authors' affiliations) and the total number of institutions participating in research.

2 Literature Review

A variety of bibliometrics studies in the last 30 years highlighted the importance of different factors (or *proxies*) of the presumed quality of research produced by researchers, institutions, and countries. In particular, they showed how researchers' performance can be affected by factors such as gender [9], location [7], reputation [18], centrality in the co-authorship network [19], online presence [20], and so on. For instance, Jadidi et al. [9] investigated gender-specific differences on about 1 million computer scientists over the course of 47 years and observed that women are on average less likely to adopt the collaboration patterns associated with a strong research impact. Petersen et al. [18] introduced an approach for quantifying the influence of an author reputation on their future research impact and found that reputation is associated with the citation count of articles, but only during the early phase of the citation lifecycle. Sarigol et al. [19] demonstrated that a classifier based only on co-authorship network centrality metrics can predict with high precision whether an article will be highly cited within five years after the publication. Thelwall et al. [20] showed that there is a significant association between eleven tested altmetrics and citations in the Web of Science dataset.

Some groundbreaking work focused in particular on the role of countries, cities, and organisations (e.g. university, research institutes) and highlighted the

[1] Springer Nature SciGraph, `https://www.springernature.com/gp/researchers/scigraph`

great discrepancy in quality and quality of the research produced by different nations. For instance, May [12] analysed the numbers of publications and citations of different countries in the 1981-1994 period using the Institute for Scientific Information database (Ed. then Thomson ISI and finally Clarivate Analytics), which included more than 8,4 million papers and 72 million citations. In accordance with our results, the authors found that the countries that produced the highest share of research papers more than 20 years ago were USA, United Kingdom, Japan, Germany, France, Canada, Italy, India, Australia, and Netherlands. King [10] built on this work and analysed the 1993-2002 period adopting again the Thomson ISI dataset. Ten years after May's study, the most important countries regarding research were essentially the same. In particular, King found that the countries that produced most of the top 1% highly cited publications were USA, United Kingdom, Germany, Japan, France, Canada, Italy, Switzerland, Netherlands, and Australia. Pan et al. [15] continued this line of work by performing a systematic analysis of citation networks between cities and countries in the 2003-2010 period. In accordance to our findings, they found that the citation distribution of countries and cities follows a power law. According to their citation rank, the main producer of research in the period under analysis were USA, United Kingdom, Germany, Japan, France, Canada, China, Italy, Netherlands, and Australia. Interestingly, they also argued that a necessary (but not sufficient) condition for a country to reach an impact larger than the world average is to invest more than about 100,000 USD per researcher annually.

Other studies are more restricted in scope and focus either on specific locations [3,1,13] or on specific research areas, such as tropical medicine [5], nanomedicine [23], biomedicine [6], and e-learning [8]. A good review of bibliometrics studies that explicitly take into account the spatial factor can be found in Frenken et al. [7]. Unlike the aforementioned analyses, in this preliminary study we (i) focused on the temporal evolution of countries and institutions in conference papers, (ii) performed an analysis on the first and last authors' affiliations, and (iii) addressed specific high-tier conferences in the domain of semantic web and digital libraries during the 2003-2017 period.

3 Data

A main premise for our study is the availability of a scholarly knowledge dataset containing information about authors' affiliations sufficiently detailed and structured, i.e. including both institution name and country, possibly disambiguated through a persistent identifier.

For the time being, given the preliminary character of this analysis, we kept intentionally out of consideration pay-walled data sources such as Scopus[2], Web of Science[3], and Microsoft Academic[4], and we focused on what can be freely retrieved on the Web, in the spirit of open and reproducible science [22].

[2] Scopus, `https://www.scopus.com`
[3] Web of Science, `https://clarivate.com/products/web-of-science`
[4] Microsoft Academic, `https://academic.microsoft.com`

Some top-quality scholarly datasets such as DBLP [11] and Semantic Scholar[5] are not apt to our study as they miss essential information about authors' affiliations. Other datasets technically provide authors' affiliations, but the relevant metadata are often incomplete. For example, Crossref[6], despite declaring a field devised for affiliations in their metadata API JSON format[7], provides in a minority of cases a simple array of affiliation strings. Besides, affiliation strings often exhibit several well-known ambiguity issues due to (i) alternate forms (e.g., "Open University" and "The Open University"), (ii) different languages (e.g., "Università di Pisa" and "University of Pisa"), (iii) different granularity and missing information (e.g., "Knowledge Media Institute, Milton Keynes").

After an analysis of current solutions for selecting a dataset curated at the source with regards to these aspects, our choice fell onto SciGraph[8], a LOD dataset published and curated by Springer Nature. To the best of our knowledge, SciGraph is the only large-scale dataset providing reconciliation of authors' affiliations by disambiguating and linking them to an external authoritative datasets in terms of institutions (in this case GRID, the Global Research Identifier Database[9]). In its entirety, SciGraph consists of 78 distinct datasets and includes about 2 billion triples describing research literature objects such as journal articles, conference papers, books, and monographs published by Springer Nature and spanning over a broad set of topics such as computer science, medicine, life sciences, chemistry, engineering, astronomy, and more.

For our analysis we focused on conferences proceedings as conferences are the focal point of networking and knowledge exchange among practitioners. To this end, we downloaded from SciGraph the *books* and *book chapters* datasets spanning from 1996 to 2017 and the *conferences* dataset linking together all the books related to the same conference series. Additionally, we downloaded the ancillary GRID dataset[10] providing a high-quality and curated database of institutions and organisations participating in research. These datasets were loaded in a graph database[11] resulting in a graph of 313,035,870 triples. Then we extracted via a SPARQL query[12] a TSV (tab-separated values) dump describing all the authors' contributions[13] to papers published in conference proceedings. This raw dataset counts 1,770,091 contributions for a total of 506,049 unique papers, accepted in 1,028 conferences.

[5] Semantic Scholar, `https://www.semanticscholar.org`
[6] Crossref, `https://www.crossref.org`
[7] `https://github.com/Crossref/rest-api-doc/blob/master/api_format.md`
[8] SciGraph datasets, `http://scigraph.springernature.com/explorer/downloads/`
[9] GRID, `https://www.grid.ac`
[10] GRID dataset, `https://www.grid.ac/downloads`
[11] GraphDB, `http://graphdb.ontotext.com`
[12] `https://github.com/andremann/SAVE-SD-2018/blob/master/extract.sparql`
[13] For the sake of clarity, if paper p is authored by authors a_1 and a_2, two distinct *contributions* (i.e. two distinct rows) are present in our dataset, one for each author.

4 Methodology

Since we intended to address both general and specific trends, we performed a *macro analysis*, on the full dataset, and a *micro analysis*, on three high-tier conferences.

In the *macro analysis* we considered all conferences in the 1996-2016 period. We did not consider 2017, since in this year we observed a fairly lower number of contributions and a large amount of unresolved affiliations. The resulting dataset includes 1,664,733 contributions (477,921 unique papers), of which 946,165 contributions are attributed to 1,016 unique conference series.

For the *micro analysis* we focused instead on three high-tier conferences in the fields of semantic web and digital libraries: the International Semantic Web Conference (ISWC), the Extended Semantic Web Conference (ESWC), and the International Conference on Theory and Practice of Digital Libraries (TPDL). We selected them for two main reasons. First, we want to perform this preliminary analysis on familiar venues near to our field of expertise. In the second instance, we were interested in comparing ISWC and ESWC, which are considered the two top conference in the semantic web domain and they traditionally tend to attract quite different demographics. The first is more international, while the second (previously called "European Semantic Web Conference") is more grounded in Europe. Focusing the analysis on three conferences enabled us to manually curate and enrich their data and therefore produce a very comprehensive representation of the involved institution and countries.

The datasets of these conferences were extracted from the raw dataset by selecting the contributions with the relevant DBLP conference series identifier (respectively *semweb*, *esws* and *ercimdl*). In some cases we deliberately chose to manually integrate some conference editions that we found missing (e.g., ISWC 2007 and 2015) and drop contributions that had been mistakenly attributed to the wrong conference (e.g., the First International Workshop of Semantic Web Services and Web Process Composition). For reasons beyond our knowledge, a conference edition appears to be missing from SciGraph (i.e., ESWC 2007) and a couple of others count less contributions than expected (i.e., TPDL 2014 and 2015). However, these few missing and circumscribed data points should not affect the overall validity of our analysis.

The manual curation phase principally aimed at resolving missing affiliations and linking them to correct institutions in the GRID database. In particular, for each contribution whose affiliation details (i.e. gridId, organisation name, city, and country) were empty, we used its affiliation string (a plain "catch-all" text field) to infer the missing pieces of information. Often, for lack of clarity of such a string, we availed of information accessible in the Springer web page about the paper and from institutional websites in order to resolve the affiliation correctly. Whenever GRID provided no entry for the institution in question, yet we were able to narrow down at least its country (e.g. aCompany GmbH), we opted for "minting" a fictional identifier. When even this was not possible, we had no other option but to leave the affiliation unresolved. Fortunately, our enrichment procedure left our datasets with a minority of unresolved contributions, as we

discuss later. We argue that this process, even if time consuming, enabled us to analyse affiliations with a good granularity and to take into account also institutions involved in a small number of research outputs. Table 1 summarises the key features about the datasets used in our analysis.

For each dataset, we took in consideration the author order and hypothesise that the first author indicates the *initiator* of a research effort, while the last author indicates the professor or the research line manager acting as an *overseer* of the work; a hypothesis that seems reasonable in many disciplines, and especially computer science. We validated this intuition by analysing the name of the researchers that appeared most as last author in the datasets under analysis. In the macro analysis dataset, we found as overseers a number of very influential scientists that lead significant research groups, such as Dinggang Shen, director of several units at UNC-Chapel Hill, Jason H. Moore, director of the Institute for Biomedical Informatics at the School of Medicine of the University of Pennsylvania, Zhang Mengjie, and so on. Similarly, in the semantic web field we encountered influential professors and scientists such as Ian Horrocks, Mark A. Musen, Stefan Decker and Sören Auer. Of course this hypothesis does not hold in all the cases (e.g. papers in which the order is alphabetical) and does not reflect a common custom for all academic disciplines (e.g. in Humanities & Social Sciences); however, we believe that this can be a good approximation that works well for this study.

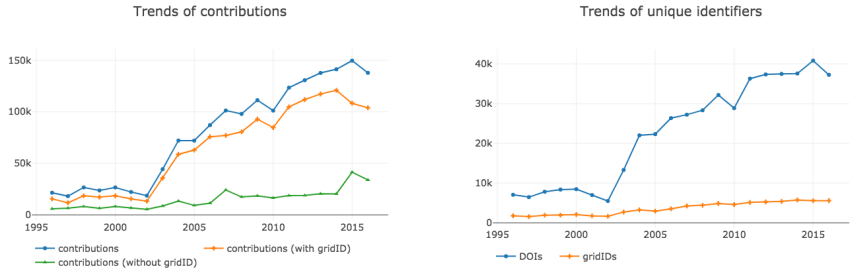
We also analysed trends about papers (identified by unique Digital Object Identifiers, *DOIs*), countries, and institutions (identified by unique *gridIDs*) over time, as well as their distributions across the entire observed period. Besides, we tried to assess to what extent the research landscape is open (or closed) to changes by measuring the variability of country rankings over the years. To this end, we defined as rate of change r_{change} the percentage of new entries (not considering permutations) entering in a top-n rankings from one year to the following. For example, if in year x the top-3 ranking is $\{a, b, c\}$ and in year $x + 1$ is $\{a, c, d\}$ then $r_{change} = 0.33$.

The result shown in the following are obtained by analysing the datasets within a Python notebook[14] availing of Pandas library[15]. For reproducibility purposes, the curated datasets and the Python notebook are accessible on Github[16]. Due to Github limitations on files size, the dataset used for the macro analysis has not been uploaded (851 MB); however, it can be easily reconstructed following the methodology we just described. All the plots here included, and many others not reported for the sake of space, are available online[17] as well. As the plots are rich in content, the images reported here cannot adequately render all the information available. Therefore, we strongly suggest the reader to consult also the online resources.

[14] Jupiter notebook, `https://ipython.org/notebook.html`
[15] Pandas library, `https://pandas.pydata.org`
[16] Code and datasets, `https://github.com/andremann/SAVE-SD-2018`
[17] `http://nbviewer.jupyter.org/github/andremann/SAVE-SD-2018/blob/master/Analysis.ipynb?flush_cache=true`

	Macro analysis	Micro analysis		
		ISWC	ESWC	TPDL
observation period	1996-2016	2003-2016	2004-2017 (excl. 2007)	2003-2017
contributions	1,664,733	3,924	4,224	3,271
unique papers (DOIs)	477,921	1,028	1,141	919
countries	163	44	54	52
institutions (gridIDs)	14,773	3,739	4,076	3,208
conference series	1,016	-	-	-

Table 1: Features of the datasets used for our analysis



(a) (b)

Fig. 1: Trends of contributions (with and without resolved affiliations), papers and institutions

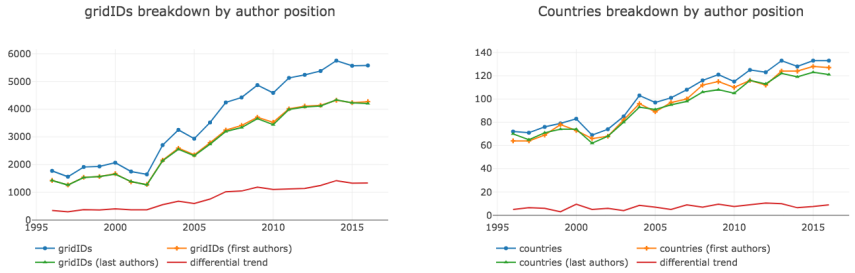
5 Results

In this Section, we report the results emerged from our macro and micro analysis. The discussion of such results can be found in Section 6.

5.1 Macro analysis

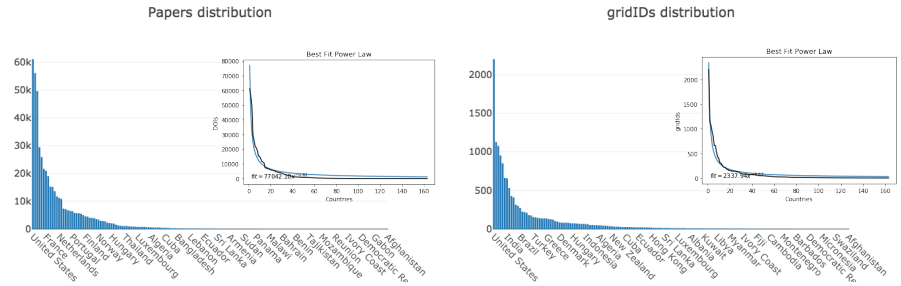
The number of contributions for each year, either with or without resolved affiliations, is reported in Figure 1a. We can notice how information about authors' affiliation is present in the majority of contributions in our dataset. Figure 1b shows the number of unique papers (*DOIs*) and the number of unique institutions (*gridIDs*) over the years. Despite a scale factor, the two trends are correlated with a Pearson correlation coefficient[17] of 0.987, suggesting that not only the volume of research literature has increased, but also that the number of institutions contributing to research has gone through the same trend.

Figure 2a presents the number of institutions involved in research over time and highlights in two dedicated series the number of institutions appearing as affiliations of the first (in yellow) and last authors (in green) respectively. For the sake of clarity, we included also the differential trend (in red) between



(a) (b)

Fig. 2: Institutions and countries breakdown according to author position



(a) (b)

Fig. 3: Papers and institutions distributions across countries and their power law approximation

first/last authors' affiliations and all the others, by computing $gridIDs_{total} - mean(gridIDs_{first}, gridIDs_{last})$. The figure suggests that there is a substantial gap between the number of institutions that initiate (first author) and overseer (last author) a research endeavour versus the total number of institutions involved in research. Also, this gap appears to grow over time despite the fact that the average number of authors per paper does not exhibit the same growth, oscillating steadily between 2.6 and 3.3 in the same time interval (not reported here for space reasons, but available online). We will investigate this phenomenon further in the micro analysis.

Similarly, Figure 2b highlights the trend of countries in function of author position. Also in this case, we see a gap between the number of first/last authors' countries of affiliation and the total number of countries involved in research. The differential trend oscillate from 5 to 9 over the observed period, despite being not remarkably growing as in the case of institutions. We believe that this is due to the naturally limited number of countries, as opposed to the virtually unbounded number of new institutions that keep appearing each year.

Figure 3a reports the distribution of papers among countries over the observed period without taking initiators and overseers into account. The distribution is heavily skewed in favour of USA, China, and Germany, highlighting a potential bias in the dataset. Indeed, a manual inspection of the dataset revealed the presence of many local Asian, Chinese, German, and American conferences. Despite the potential bias, the power law characteristic of the distribution is evident. In the figure inset, we report the best fit power law obtained by fitting the data points with the least squares method to a power law function of the type $y = ax^s$, with $s < 0$. Interestingly, the power law characteristic of the paper distribution over countries is also valid in each year across the period. We verified this by checking Pareto rule[16] for every year, and discovered that invariably 20% of the countries produces more than 80% of the papers.

The distribution of institutions over countries (i.e. the number of institutions present in a given country) follows as well a power law, as shown by Figure 3b. For the sake of space, we omitted the details about the distributions of papers for first and last authors, which the reader can consult online.

We also noticed that the average r_{change} for the top-5, top-10 and top-25 across the observed period yielded 0.13, 0.09, and 0.08 respectively. This suggests that (i) year by year it is fairly hard for outsiders to break in a top-n, and (ii) that it gets harder and harder as the top-n set broadens. In addition, over the 21 year span of our observation, the top-5 has been visited by 10 countries, the top-10 by 16 and the top-25 by 36. For example, the top-10 has been visited by USA (21), Germany (21), Japan (21), United Kingdom (21), Italy (21), France (21), Spain (19), Canada (16), China(13), Netherlands (9), South Korea (6), India (6), Poland (5), Russia (4), Australia (3), Switzerland (3); further details are available online.

5.2 Micro analysis

Here we summarise the results obtained by analysing the three high-tier conferences (i.e., ISWC, ESWC and TPDL).

Figure 4 and Figure 5 show respectively the number of contributions, and the number of papers and institutions contributing to the conferences over the years. Since we manually curated the three datasets, the percentage of unresolved affiliations is much lower than the one of the macro analysis. Again we can observe a high correlation between the number of papers accepted and the number of contributing institutions. As opposed to what we observed in the macro analysis, this time the number of papers and institutions are within the same order of magnitude. This can be explained considering that the number of papers accepted each year by a conference is naturally limited, whereas there is not limitation to the number of institutions that can apply.

Similar to what was observed in the macro analysis, Figure 6 shows the number of institutions contributing to the conferences and highlights the trends of the ones appearing as first and last authors' affiliations. As in the previous analysis, the growing gap between the institutions associated with first/last authors

(a) (b)

(c)

Fig. 4: Trends of contributions with and without resolved affiliations

and the total number of affiliations is present for all the three conferences as suggested by the differential trend.

We investigated further and retrieved the sets of institutions never appearing as either first or last authors' affiliations throughout the entire observed periods (available online). Here it can be noted how prestigious universities and research centres appear side by side with smaller firms and less well-known universities or institutions. This result indicates that the gap is "populated" by institutions that at some point collaborated in semantic web research (or digital libraries) making it through, whereas they never stand out on their own (for reasons beyond our knowledge) in the communities of the respective conferences. Institutions like national libraries, the European Bioinformatics Institute, the British Geological Survey, the National Institute of Standards and Technology, and so on, provided interesting research case studies or support that eventually culminated in a publication, but apparently never happened to author a paper on their own. We also verified that the intersection between these sets across different conferences is not empty, suggesting that a few institutions struggled to surface as key contributors, despite being present in either community.

It is important to stress that the sets of institutions appearing as first/last authors' affiliation in different years are very likely to differ; it is not the intention

Fig. 5: Trends of papers and institutions

of this study to suggest that institutions initiating or overseeing research are essentially unaltered throughout time.

Figure 7 shows the trend of countries contributing to the conferences, highlighting country affiliations of first and last authors. Consistently to what we observed in the macro analysis a gap is present and growing.

Figure 8 and Figure 9 again confirm the results shown in previous section even at micro level: the distribution of papers and institutions across countries indeed follows a power law. However, the power law characteristic surfaces only across the entire observed period as, in general, in a single year the Pareto rule might not be verified mainly because of insufficient data points (i.e. in a single conference edition the number of papers is limited). In this case, evaluating a top-n stratified rate of change for single conferences gets difficult as the set of countries participating in a single year can be quite limited. However, as can be seen in the results online, the situation in the top-10 achieves an average $r_{change} \approx 0.23$. Moreover, it appears that the top-10 is regularly visited by a small number of countries. In particular, in ISWC only 13 countries enter the top-10 more then 3 times in the 14 year period. Similarly, only 14 countries enter the top-10 in ESWC and TPDL.

Finally, we noticed a stronger presence of European countries in ESWC than in the other two conferences; this is probably due to the initial local target of

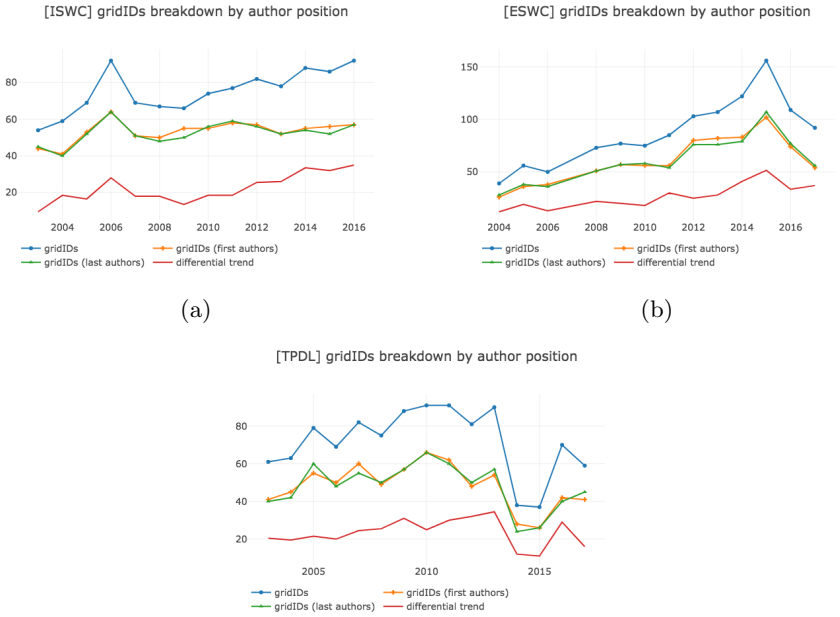


Fig. 6: Institutions breakdown according to author position

the conference. China is quite involved in the semantic web community, but, perhaps surprising, is less active in the TPDL conference and never appears in the top-10.

6 Discussion

The study of authors' affiliations in research literature has been performed for decades as it can provide answers to socioeconomic questions and frame academic research on a geopolitical canvas rather than studying it as an isolated complex system. In this work we analysed four datasets distilled from Springer Nature Scigraph and provided results on both a macro and a micro scale, focusing on three different high-tier conferences.

The results, in accordance with previous studies [12,10,15], showed that distributions of papers and institutions across countries still exhibit a power law characteristic in the period 1996-2016. In addition, our analysis of the turnover rate highlights that not only top ranks in research are limited to a handful of countries and institutions, but that the situation appears also to be stagnant towards the lower ranks. In general, this reflects the intuition that well-formed research communities exhibit a sort of resistance towards the permeation of outsiders not always sharing knowledge and best practices consolidated over the

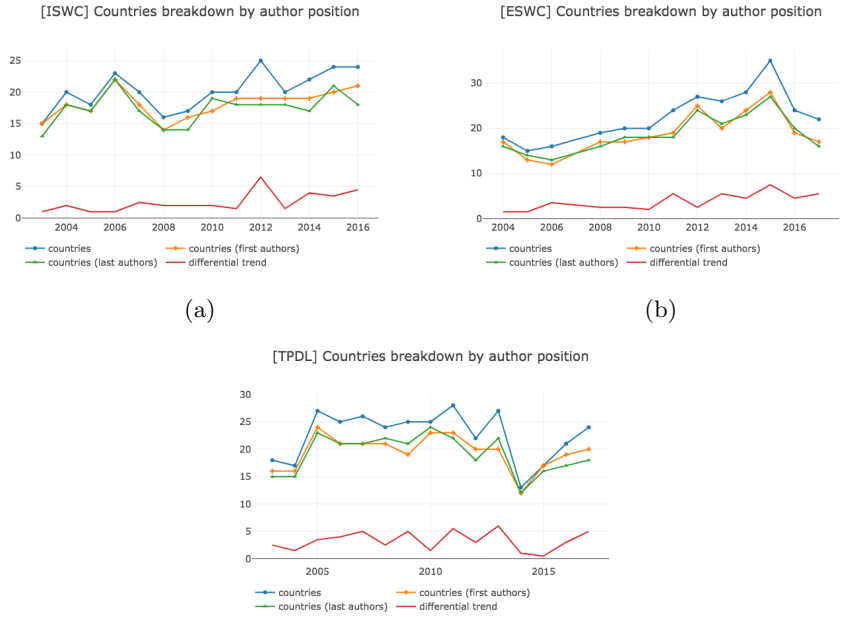


Fig. 7: Countries breakdown according to author position

years. Therefore, we believe that this phenomenon is worth studying further. Besides, the papers eventually accepted in conferences is a minimal fraction of the whole amount of submissions; a much clearer view about openness/closeness of conferences and research communities could be achieved by having access to data about rejected papers held in conference management systems such as EasyChair[18] or ConfTool[19].

The results from our study on first and last authors' affiliations show that, in principle, weighting authors' contributions is an intuition that can provide different keys to interpret data. Other studies dealing with researchers' seniority, for example, take into account the volume of publications produced by a single author throughout a sliding window of W years [21], or the number of consecutive years of publishing activity [9]. We intend to further investigate these techniques and test further our intuition in order to understand its applicability in other disciplines and extend the approach by including other metrics (e.g., seniority); nonetheless, the preliminary results are indeed interesting.

Furthermore, a final remark has to be spent about the very peculiar nature of the data here considered: conference papers; usually not covered by traditional scientometrics and bibliometrics studies that instead mainly focus on journals.

[18] EasyChair conference management system, `http://easychair.org`
[19] ConfTool conference & event management software, `http://www.conftool.net`

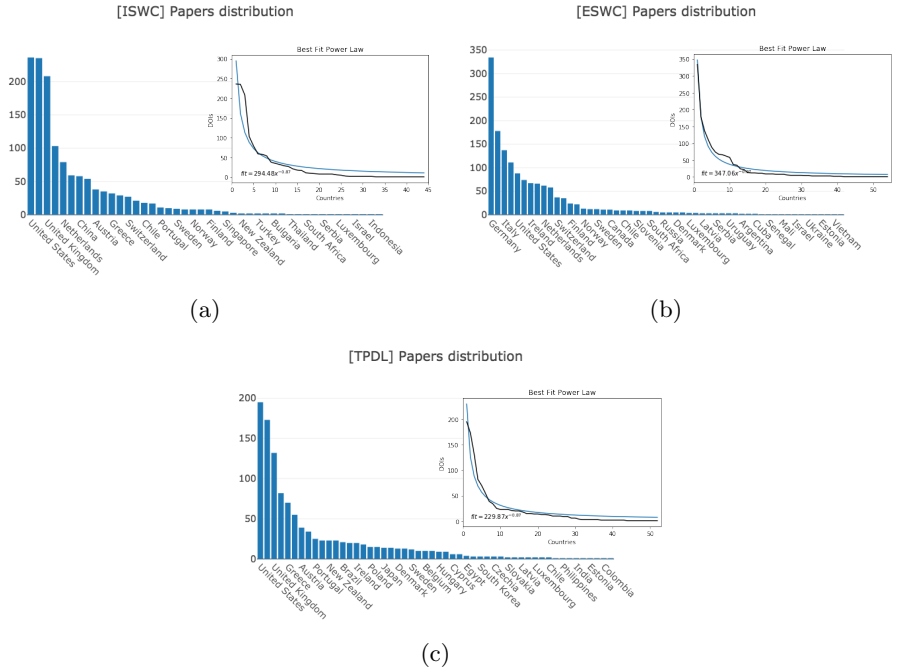


Fig. 8: Paper distributions across countries and power law approximations

Unlike journal papers, having a publication accepted in conference proceedings often requires that at least an author is registered to the event and presents the work at the venue. This aspect has major implications that need to be studied further. For example, scientists' mobility is subject to economic and geopolitical factors such as geographic distance, budget availability for travels, and travel bans. In some cases, being physically present at the conference venue means taking long-haul flights; for some countries, such as Australia and similarly rather isolated countries, the chances of being poorly connected to the conference venues are high. In other cases, despite feasible connections are available, the physical attendance might be hindered by economic factors, that in turn can depend on strategic and political decisions within the single country. Finally, factors driven by international politics can play a major role too. In several occasions, travel bans disrupted scientists' mobility; in 2013, for example, NASA prevented Chinese nationals to set foot in the space agency's Ames research centre in California[20]. Furthermore, citizens of countries with an important Muslim background always have encountered more difficulties for getting travel visas to European or USA countries. In addition, recent USA's international policies and travel restrictions, possibly have made this even worse [14]. However, it has to be noted

[20] `https://www.theguardian.com/science/2013/oct/05/us-scientists-boycott-nasa-china-ban`

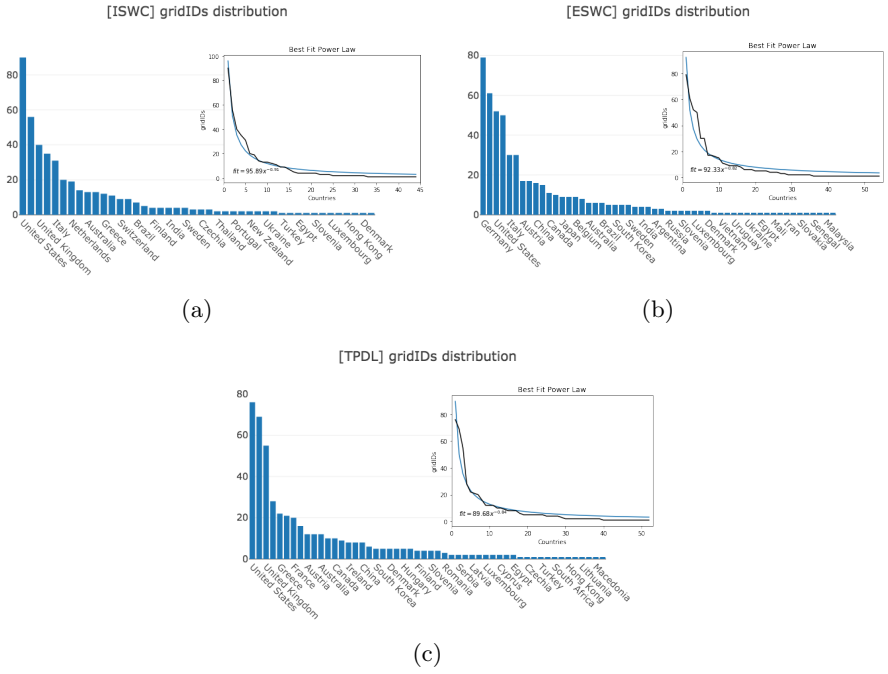


Fig. 9: gridIDs distributions across countries and power law approximations

that these concerns about researchers' freedom of movement affect only conference papers in which all the authors are subject to the same kind of restrictions; in the case of papers whose authors have heterogeneous affiliations, for example, the author with less restrictive constraints is, in principle, free to reach the venue and present the findings on behalf of the colleagues. All these implications are worth studying. To this end, a future extension of this work could include the comparison of country rankings among high-tier conferences and journals from a controlled set of academic fields in order to analyse whether the freedom of mobility has an impact or not on how countries perform.

In conclusion, we advocate openness and transparency for research literature metadata. It is detrimental to research itself to relinquish information about venues, papers, authorship and much more in data silos hard (or almost impossible) to access. Datasets like SciGraph are a bless for researchers working on scholarly analytics and such initiatives should be fostered. Moreover, new best practices for declaring unambiguous authors' affiliations should be devised in order to facilitate the work of researcher working with scholarly knowledge. Being able to access high quality research literature metadata is key for enabling large-scale analytics and cross-correlate scholarly knowledge with external datasets and hopefully get better and more thorough insight on the existing global dynamics prevailing in academic research.

Acknowledgements

We would like to thank the SciGraph team, especially Dr. Michele Pasin, whose work and prompt response made this study possible.

References

1. Börner, K., Penumarthy, S.: Spatio-temporal information production and consumption of major us research institutions. Proceedings of ISSI Volume 1 (2005)
2. Bornmann, L., Mutz, R.: Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. Journal of the Association for Information Science and Technology 66(11), 2215–2222 (nov 2015)
3. Carvalho, R., Batty, M.: The geography of scientific productivity: Scaling in US computer science. Journal of Statistical Mechanics: Theory and Experiment (10) (2006)
4. Egghe, L.: Power laws in the information production process: Lotkaian informetrics. Emerald Group Publishing Limited (2005)
5. Falagas, M.E., Karavasiou, A.I., Bliziotis, I.A.: A bibliometric analysis of global trends of research productivity in tropical medicine. Acta tropica 99(2-3), 155–159 (2006)
6. Falagas, M.E., Michalopoulos, A.S., Bliziotis, I.A., Soteriades, E.S.: A bibliometric analysis by geographic area of published research in several biomedical fields, 1995–2003. Canadian Medical Association Journal 175(11), 1389–1390 (2006)
7. Frenken, K., Hardeman, S., Hoekman, J.: Spatial scientometrics: Towards a cumulative research program. Journal of Informetrics 3(3), 222–232 (2009)
8. Hung, J.l.: Trends of e-learning research from 2000 to 2008: Use of text mining and bibliometrics. British Journal of Educational Technology 43(1), 5–16 (2012)
9. Jadidi, M., Karimi, F., Lietz, H., Wagner, C.: Gender Disparities in Science? Dropout, Productivity, Collaborations and Success of Male and Female Computer Scientists. Advances in Complex Systems p. 1750011 (nov 2017)
10. King, D.A.: The scientific impact of nations (2004)
11. Ley, M.: DBLP: some lessons learned. Proceedings of the VLDB Endowment 2(2), 1493–1500 (aug 2009)
12. May, R.M.: The scientific wealth of nations. Science 275(5301), 793–796 (1997)
13. Monroe-White, T., Woodson, T.S.: Inequalities in scholarly knowledge: Public value failures and their impact on global science. African Journal of Science, Technology, Innovation and Development 8(2), 178–186 (2016)
14. Morello, L., Reardon, S., Others: Scientists struggle with Trump immigration ban. Nature 542(7639), 13–14 (2017)
15. Pan, R.K., Kaski, K., Fortunato, S.: World citation and collaboration networks: Uncovering the role of geography in science. Scientific Reports 2 (2012)
16. Pareto, V., Page, A.N.: Translation of Manuale di economia politica (Manual of political economy). AM Kelley (1971)
17. Pearson, K.: Mathematical contributions to the theory of evolution. III. Regression, heredity, and panmixia. Philosophical Transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character 187, 253–318 (1896)

18. Petersen, A.M., Fortunato, S., Pan, R.K., Kaski, K., Penner, O., Rungi, A., Riccaboni, M., Stanley, H.E., Pammolli, F.: Reputation and impact in academic careers. Proceedings of the National Academy of Sciences 111(43), 15316–15321 (2014)
19. Sarigöl, E., Pfitzner, R., Scholtes, I., Garas, A., Schweitzer, F.: Predicting scientific success based on coauthorship networks. EPJ Data Science 3(1), 9 (2014)
20. Thelwall, M., Haustein, S., Larivière, V., Sugimoto, C.R.: Do altmetrics work? Twitter and ten other social web services. PloS one 8(5), e64841 (2013)
21. Verleysen, F.T., Weeren, A.: Clustering by publication patterns of senior authors in the social sciences and humanities. Journal of Informetrics 10(1), 254–272 (2016)
22. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., Others: The FAIR Guiding Principles for scientific data management and stewardship. Scientific data 3 (2016)
23. Woodson, T.S.: Research Inequality in Nanomedicine. Journal of Business Chemistry 9(3), 133–146 (2012)